# Score-Based Diffusion Priors for Adaptive Conformal Inference under Distribution Shift

**Xiangyu Jiang**[1]

[1]Southwestern University of Finance and Economics, Chengdu, China

## Abstract

Conformal prediction provides a distribution-free coverage guarantee for predictive inference, yet its validity degrades under distribution shift, a common challenge in real-world deployment. We introduce DiffConf, a framework that uses score-based diffusion models as expressive priors over the data-generating process to enable adaptive conformal inference under temporal and covariate distribution shifts. The main idea is that the score function learned by a diffusion model encodes rich geometric information about the data manifold, which can be repurposed to construct nonconformity scores that are sensitive to distributional changes. We derive a diffusion-guided conformity score that integrates the learned score field with a lightweight online recalibration mechanism, providing finite-sample marginal coverage guarantees even when the data distribution evolves over time. Theoretically, we establish that DiffConf achieves asymptotic conditional coverage under mild regularity conditions on the drift rate, and we prove a regret bound that scales gracefully with the complexity of the distribution shift. Experiments on synthetic benchmarks, real-world tabular regression tasks, and high-dimensional image datasets demonstrate that DiffConf produces prediction sets that are simultaneously valid and more efficient than existing adaptive conformal methods, reducing average set size by 6–39% across the real-data benchmarks (and by over 50% under large synthetic mean shifts) while keeping coverage within one point of target.

## 1 INTRODUCTION

Reliable uncertainty quantification is central to trustworthy machine learning. As predictive models are deployed in safety-critical domains, including medical diagnosis [Gal and Ghahramani, 2016] and autonomous driving [Lakshminarayanan et al., 2017], the ability to communicate *when* and *how much* predictions can be trusted becomes as important as the predictions themselves. Conformal prediction [Vovk et al., 2005] provides a principled framework for this purpose, offering finite-sample coverage guarantees under the mild assumption of exchangeability.

However, exchangeability is routinely violated in practice. Distributions shift over time due to changing environments, evolving populations, or feedback loops [Barber et al., 2023], causing standard conformal prediction to become miscalibrated. This has motivated *adaptive* conformal inference [Gibbs and Candès, 2021, Angelopoulos et al., 2024], which maintains valid coverage under distribution shift by dynamically adjusting a scalar threshold based on observed coverage errors. While effective in mild shift regimes, these methods treat the nonconformity score as a fixed black-box and do not exploit structural knowledge about *how* the distribution is changing. When shifts are rapid or complex, threshold-only adaptation reacts only through the delayed scalar coverage residual: on the synthetic mean shift of Figure 2, threshold-only baselines need several hundred steps to re-attain the target after an abrupt shift, during which intervals are miscalibrated or inflated.

We take a different approach: rather than adapting only the threshold, we adapt the *nonconformity score itself* using the geometric structure learned by a score-based diffusion model [Song et al., 2021, Ho et al., 2020]. Diffusion models approximate the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ across a continuum of noise levels, encoding a multi-scale representation of the data distribution. This score field provides a natural measure of how "typical" a data point is relative to the learned distribution, and in particular, how this typicality changes as the distribution shifts. At coarse noise levels, the

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:2517–2537.

score captures global distributional changes; at fine levels, it is sensitive to local structural changes such as the emergence of new subpopulations.

We introduce **DiffConf** (**Diff**usion-guided **Conf**ormal inference), which integrates score-based diffusion priors with adaptive conformal prediction. The core is a *diffusion-guided conformity score* that combines a base model's residuals with a score-based density contrast term, fed into an online recalibration procedure that adjusts both the score function parameters and the quantile threshold simultaneously. DiffConf does not require the diffusion model to generate high-quality samples; it only requires that the learned score function captures the relevant distributional geometry.

Our contributions are as follows:

- We propose DiffConf, a framework that uses score-based diffusion models to construct distribution-shift-aware nonconformity scores for adaptive conformal inference. While recent work has explored generative models for uncertainty estimation [Jazbec et al., 2025], DiffConf is, to the best of our knowledge, among the first to repurpose the *score function* of a diffusion model specifically for conformal prediction under distribution shift.
- We derive a diffusion-guided conformity score that integrates the learned score field at multiple noise levels, and prove that it satisfies a key monotonicity property with respect to distributional distance (Proposition 1).
- We establish finite-sample marginal coverage guarantees for DiffConf under arbitrary distribution shift (Theorem 1), prove asymptotic conditional coverage under regularity conditions on the drift rate (Theorem 2), and derive an efficiency regret bound that improves upon existing adaptive conformal methods when the diffusion prior is informative (Theorem 3).
- We evaluate DiffConf on synthetic, tabular, and image datasets, demonstrating that it achieves near-target coverage while producing prediction sets that are 6–39% smaller than coverage-maintaining adaptive conformal baselines, with the largest gains on streams where the shift displaces the residual distribution (rather than merely rescaling it), and an advantage that grows with the displacement magnitude.

# 2 BACKGROUND AND RELATED WORK

## 2.1 CONFORMAL PREDICTION

Conformal prediction [Vovk et al., 2005, Lei et al., 2018] constructs prediction sets $\mathcal{C}(X_{n+1}) \subseteq \mathcal{Y}$ satisfying $\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$ for a user-specified miscoverage level $\alpha$. Given a nonconformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and calibration scores $s_i = s(X_i, Y_i)$, the set is $\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \hat{q}\}$, where $\hat{q}$ is the appropriate quantile of $s_1, \ldots, s_n$. This guarantee requires only exchangeability [Angelopoulos and Bates, 2023]. Conformalized quantile regression (CQR) [Romano et al., 2019] combines quantile regression [Koenker and Bassett, 1978] with conformal calibration for adaptive intervals, extended by Chernozhukov et al. [2021] for conditional validity.

## 2.2 ADAPTIVE CONFORMAL INFERENCE UNDER DISTRIBUTION SHIFT

When exchangeability is violated, standard conformal prediction fails. Barber et al. [2023] introduced weighted conformal prediction using likelihood ratios, but density ratio estimation is challenging in high dimensions. Gibbs and Candès [2021] proposed Adaptive Conformal Inference (ACI), which adjusts the miscoverage level online:

$$\alpha_{t+1} = \alpha_t + \gamma\big(\alpha - \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\}\big), \tag{1}$$

guaranteeing long-run coverage regardless of the data-generating process. Extensions include decaying step sizes [Angelopoulos et al., 2024], regret bounds [Gibbs and Candès, 2024], parameter-free variants [Podkopaev et al., 2024], and covariate/label shift corrections [Tibshirani et al., 2019, Podkopaev and Ramdas, 2021].

All these methods adapt only the *threshold* while keeping the nonconformity score fixed. When shifts affect the data manifold geometry, a fixed score assigns systematically biased conformity values, leading to inefficient prediction sets. Our work addresses this by adapting the score function itself via a diffusion model.

## 2.3 SCORE-BASED DIFFUSION MODELS

Score-based diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song et al., 2021] define a forward noising process that transforms data into noise, and learn to reverse this process. Building on classical stochastic analysis [Anderson, 1982], the forward process is an SDE:

$$\mathrm{d}\mathbf{x}_t = f(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t, \quad t \in [0, T], \tag{2}$$

where $f$ is the drift, $g$ the diffusion coefficient, and $\mathbf{w}_t$ a Wiener process. A neural network $s_{\boldsymbol{\theta}}(\mathbf{x}, t)$ approximates the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ [Hyvärinen, 2005] via denoising score matching [Song and Ermon, 2019, Ho et al., 2020]:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}\big[\|s_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\|^2\big]. \tag{3}$$

The learned score captures rich geometric information about the data manifold [Karras et al., 2022]. Diffusion models have been applied to inverse problems [Chung et al., 2023], Bayesian inference [Bao et al., 2022], and uncertainty estimation [Jazbec et al., 2025], with significant improvements

in sample quality [Nichol and Dhariwal, 2021, Dhariwal and Nichol, 2021, Rombach et al., 2022]. However, their potential for conformal prediction has remained unexplored.

**Uncertainty quantification in deep learning.** The broader UQ literature includes Bayesian approaches such as MC Dropout [Gal and Ghahramani, 2016], deep ensembles [Lakshminarayanan et al., 2017], and variational inference [Blundell et al., 2015, Kingma and Welling, 2014], as well as calibration methods [Gneiting and Raftery, 2007, Kuleshov et al., 2018]. While useful, these lack the distribution-free coverage guarantees of conformal prediction. Nalisnick et al. [2019] showed that even deep generative models can fail to detect out-of-distribution inputs, complementing earlier observations [Hendrycks and Gimpel, 2017] and highlighting the need for distribution-free approaches. Recent work has sought to combine Bayesian and conformal methods [Wilson and Izmailov, 2020], but combining generative model priors with conformal inference remains an open problem.

# 3 METHOD: DIFFCONF

We present DiffConf, our framework for adaptive conformal inference guided by score-based diffusion priors (Figure 1).

## 3.1 PROBLEM SETTING

At each time step $t$, we observe $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and must produce a prediction set $\mathcal{C}_t(X_t) \subseteq \mathcal{Y}$ before observing $Y_t \in \mathcal{Y} \subseteq \mathbb{R}$, where $(X_t, Y_t) \sim P_t$ for a time-varying distribution $P_t$. Our goal is long-run coverage $\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\} \leq \alpha + o(1)$ while minimizing average set size. We assume access to a pre-trained base predictor $\hat{f}$, a score network $s_{\boldsymbol{\theta}}$ trained on the *covariate marginal* $P_0^X$ of a reference distribution $P_0$ (i.e., the diffusion model operates on the input space $\mathcal{X}$, not the joint $\mathcal{X} \times \mathcal{Y}$), and a sliding window of $W$ recent observations for online calibration. Training on $P_0^X$ rather than the joint is a deliberate design choice: the score discrepancy is intended to detect covariate shift, while the residual term in (4) captures changes in the conditional $P_t(Y|X)$.

**$P_0$ as a Bayesian-style prior.** The reference distribution plays a role analogous to a prior, but over the *geometry of the input space* rather than over parameters or hypotheses; related bridges between Bayesian modeling and conformal calibration appear in Fong and Holmes [2021], Cruz Cabezas et al. [2025], Bariletto et al. [2025]. The score field of $P_0^X$ encodes what typical inputs look like at every noise scale and enters the prediction set only multiplicatively, through the modulation in (4); when no shift occurs, the population discrepancy vanishes (Proposition 1) and DiffConf reduces to its threshold-only backbone, so the prior is non-informative on $P_0$ itself. The EMA reference (6) is the closest analog of a posterior: a convex-combination belief update over the score field, not a likelihood-weighted Bayes rule. We claim no Bayesian credible-interval semantics; all coverage guarantees in Section 4 are frequentist.

## 3.2 DIFFUSION-GUIDED CONFORMITY SCORE

The central component of DiffConf is a conformity score that uses the diffusion model's score function to detect and adapt to distribution shift. We define the *diffusion-guided conformity score* as:

$$s_t(x, y) = \underbrace{|y - \hat{f}(x)|}_{\text{residual term}} \cdot \underbrace{\exp\big(\lambda \cdot \Delta_{\boldsymbol{\theta}}(x, t)\big)}_{\text{diffusion modulation}}, \tag{4}$$

where $\lambda > 0$ is a modulation strength parameter and $\Delta_{\boldsymbol{\theta}}(x, t)$ is the *score discrepancy function* defined as:

$$\Delta_{\boldsymbol{\theta}}(x, t) = \int_0^T w(\tau) \big\| s_{\boldsymbol{\theta}}(x_\tau, \tau) - s_{\boldsymbol{\theta}}^{(\text{ref})}(x_\tau, \tau) \big\|^2 \, \mathrm{d}\tau. \tag{5}$$

Here, $x_\tau$ denotes the noised version of $x$ at diffusion time $\tau$, and $w(\tau)$ controls the contribution of different noise levels. The integral is approximated via MC sampling over $\{\tau_1, \ldots, \tau_K\}$. The reference term $s_{\boldsymbol{\theta}}^{(\text{ref})}$ is an exponential moving average (EMA) of the network's outputs on recent data:

$$\bar{s}_t^{(\text{ref})}(\cdot, \tau) = (1 - \rho)\, \bar{s}_{t-1}^{(\text{ref})}(\cdot, \tau) + \rho\, s_{\boldsymbol{\theta}}(\cdot, \tau), \tag{6}$$

with $\rho = 0.01$. Although both terms use the same frozen network $s_{\boldsymbol{\theta}}$ (trained on $P_0^X$), they are evaluated on different inputs: the current term on the fresh sample $x_t$, and the EMA on a running average reflecting the recent past. The discrepancy thus captures how the network's output on fresh data differs from its recent behavior (see Appendix B for a detailed discussion of the theory–implementation connection).

**Proposition 1** (Monotonicity of Score Discrepancy)**.** *Let $p$ and $q$ be two distributions on $\mathcal{X}$ with finite Fisher information, and let $p_\tau$, $q_\tau$ denote their marginals under the forward SDE at diffusion time $\tau$. Define the population-level score discrepancy*

$$\bar{\Delta}(p, q) = \int_0^T w(\tau)\, \mathbb{E}_{x \sim p_\tau} \big[ \|\nabla_x \log p_\tau(x) - \nabla_x \log q_\tau(x)\|^2 \big] \, \mathrm{d}\tau. \tag{7}$$

*Then $\bar{\Delta}(p, q) \geq 0$ with equality iff $p = q$ a.e. For uniform weights $w(\tau) = 1/T$ under the VP-SDE with schedule bound $\beta_{\max}$,*

$$\bar{\Delta}(p, q) \geq \frac{1 - e^{-2\beta_{\max} T}}{2\beta_{\max} T} \cdot D_F(p \| q), \tag{8}$$

*where $D_F$ denotes the Fisher divergence.*

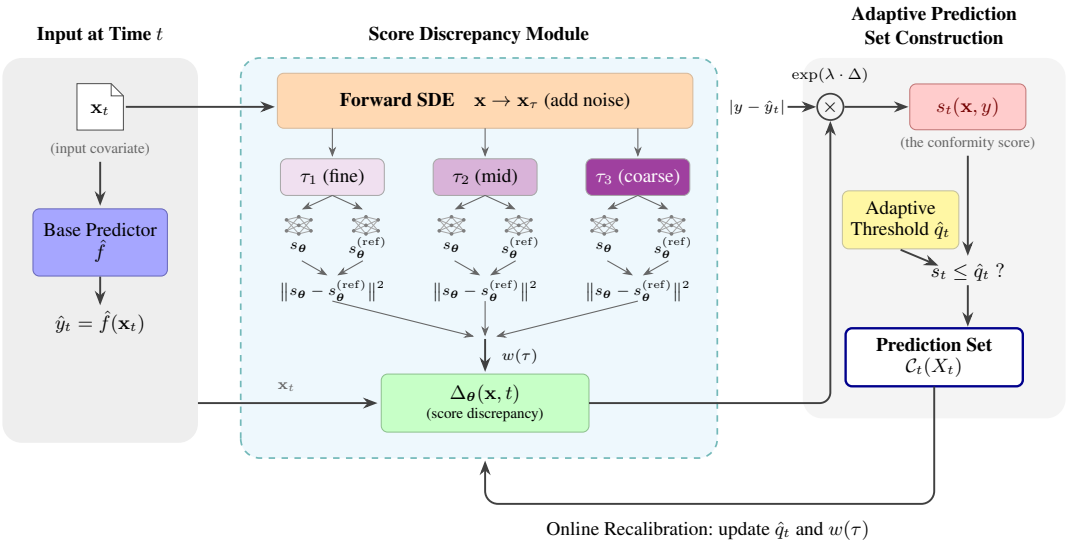


Figure 1: Overview of DiffConf. The base predictor produces $\hat{y}_t$ while the Score Discrepancy Module computes $\Delta_{\boldsymbol{\theta}}(x, t)$ across noise levels. The residual and exponentiated discrepancy form the conformity score $s_t(x, y)$, compared against threshold $\hat{q}_t$ to construct $\mathcal{C}_t(X_t)$. An online loop updates $\hat{q}_t$ and weights $w(\tau)$ after observing $Y_t$.

The proof (Appendix A) uses the exponential decay of Fisher divergence under the VP-SDE semigroup. The gap between this population-level result and the practical EMA-based estimator is accounted for by the $\epsilon_s$ terms in Theorems 2–3 (see Remark 2 in Appendix B).

**Multi-scale score discrepancy.** The weighting function $w(\tau)$ captures distributional changes at different scales: small $\tau$ is sensitive to local structure, large $\tau$ to global properties. We use a learned weighting:

$$w(\tau) = \frac{\exp(\beta_\tau)}{\sum_{k=1}^{K} \exp(\beta_{\tau_k})}, \quad \beta_\tau \in \mathbb{R}, \tag{9}$$

where the weights $\{\beta_{\tau_k}\}$ are optimized online to minimize the prediction set size subject to coverage constraints (see Section 3.3).

**Efficient computation.** We introduce two optimizations: (i) *amortized score evaluation*, caching the reference score statistics $\bar{s}_t^{(\mathrm{ref})}(x_\tau, \tau)$ so that at test time only the current-sample forward pass is needed per noise level (the reference term is a cached running average, not a fresh network evaluation); and (ii) *progressive refinement*, starting with coarse noise levels and refining only where the discrepancy exceeds a threshold, skipping fine levels when the coarse signal is small.

### 3.3 ONLINE RECALIBRATION WITH DIFFUSION GUIDANCE

The prediction set at time $t$ is:

$$\mathcal{C}_t(X_t) = \big\{y \in \mathcal{Y} : s_t(X_t, y) \le \hat{q}_t\big\}, \tag{10}$$

which for regression reduces to $[\hat{f}(X_t) \pm r_t(X_t)]$ with adaptive radius $r_t(X_t) = \hat{q}_t \cdot \exp(-\lambda \cdot \Delta_{\boldsymbol{\theta}}(X_t, t))$. The threshold is updated via a modified ACI rule [Gibbs and Candès, 2021]:

$$\hat{q}_{t+1} = \hat{q}_t + \gamma_t\big(\alpha - \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\}\big) - \eta_t \cdot \nabla_{\hat{q}} \mathcal{L}_{\text{eff}}(\hat{q}_t), \tag{11}$$

where $\gamma_t$ is the coverage step size, $\eta_t$ is the efficiency step size, and $\mathcal{L}_{\text{eff}}$ is an efficiency loss that penalizes large prediction sets:

$$\mathcal{L}_{\text{eff}}(\hat{q}_t) = \frac{1}{W} \sum_{j=t-W}^{t-1} |\mathcal{C}_j(X_j)| \cdot \mathbf{1}\{Y_j \in \mathcal{C}_j(X_j)\}. \tag{12}$$

**Two-sided center–width tracking.** The radius form above adapts only the interval *width*; when the shift also moves the center of the residual distribution (e.g., a conditional-mean drift), a symmetric interval must inflate both sides to cover an asymmetric error. We therefore generalize (11) to track the two conditional quantiles of the *signed*, locally rescaled residual $z_t = (Y_t - \hat{f}(X_t))/m_t(X_t)$, where $m_t(x) > 0$ is a local scale (the diffusion modulation $m_t(x) = \exp(-\lambda \Delta_{\boldsymbol{\theta}}(x, t))$, or a heteroscedasticity estimate $\hat{\sigma}(x)$ built from the multi-scale score features or from local statistics of the base predictor's outputs):

$$\begin{aligned} \hat{u}_{t+1} &= \hat{u}_t + \gamma_t\big(\mathbf{1}\{z_t > \hat{u}_t\} - \tfrac{\alpha}{2}\big), \\ \hat{l}_{t+1} &= \hat{l}_t - \gamma_t\big(\mathbf{1}\{z_t < \hat{l}_t\} - \tfrac{\alpha}{2}\big), \end{aligned} \tag{13}$$

yielding the (possibly asymmetric) prediction interval $\mathcal{C}_t(X_t) = [\hat{f}(X_t) + m_t(X_t)\,\hat{l}_t,\ \hat{f}(X_t) + m_t(X_t)\,\hat{u}_t]$. Each line of (13) is an online quantile update of exactly the same form as (11) applied to one tail at level $\alpha/2$, so the marginal

**Algorithm 1** DiffConf: Diffusion-Guided Adaptive Conformal Inference

**Require:** Base predictor $\hat{f}$, score network $s_{\boldsymbol{\theta}}$, noise levels $\{\tau_k\}_{k=1}^{K}$, target level $\alpha$, step sizes $\gamma_t, \eta_t$, window size $W$, modulation $\lambda$

Initialize $\hat{q}_1$, weights $\{\beta_{\tau_k}\}$, reference scores
**for** $t = 1, 2, \ldots$ **do**
  Observe covariate $X_t$
  Compute score discrepancy $\Delta_{\boldsymbol{\theta}}(X_t, t)$ via (5)
  Construct prediction set $\mathcal{C}_t(X_t)$ via (10)
  Output $\mathcal{C}_t(X_t)$
  Observe response $Y_t$
  Update threshold $\hat{q}_{t+1}$ via (11), or the pair $(\hat{l}_{t+1}, \hat{u}_{t+1})$ via (13)
  Update weights $\beta_{\tau_k} \leftarrow \beta_{\tau_k} - \eta_t \nabla_{\beta_{\tau_k}} \mathcal{L}_{\text{eff}}$
  **if** buffer full (every $W$ steps) **and** fine-tuning enabled **then**
    Fine-tune $s_{\boldsymbol{\theta}}$ on recent $W$ samples ▷ *optional; disabled in all experiments*
  **end if**
**end for**

coverage guarantee of Theorem 1 applies verbatim to each side (and hence to the interval, by a union of the two tail miscoverage bounds); when the residual distribution is symmetric the two updates coincide with the radius form. The choice between the radius form (11) and the two-sided form (13), as well as the scale instantiation $m_t$, is made once on the validation stream, never on test data.

The noise-level weights $\{\beta_{\tau_k}\}$ are updated simultaneously via online gradient descent on the same efficiency objective, subject to maintaining the coverage constraint. The complete procedure is summarized in Algorithm 1.

**Practical considerations.** DiffConf is agnostic to the diffusion architecture; in all our experiments a small MLP score network suffices, operating on raw inputs for tabular data and on frozen backbone feature embeddings for image data. The reference score estimation via exponential moving average (Section 3) adapts naturally to gradual shifts while remaining computationally lightweight. All configuration choices (the modulation strength $\lambda$, the tracking step sizes, and the radius-form vs. two-sided form of the online update) are selected on a held-out validation stream; $\lambda \in [0.1, 1.0]$ works well across datasets. Algorithm 1 includes an optional score network fine-tuning step; in all our experiments, we *disable* fine-tuning and rely solely on the online weight and threshold updates, which suffice for the shift magnitudes we consider. Fine-tuning may benefit scenarios with very large distributional drift but introduces additional computational cost and potential instability. A twelve-line minimal implementation of the online loop is given in Appendix C.6; further details are in Appendix C.

# 4 THEORETICAL ANALYSIS

We establish the theoretical properties of DiffConf. All proofs are deferred to the supplementary material.

**Assumption 1.** *The score discrepancy is bounded:* $\Delta_{\boldsymbol{\theta}}(x, t) \in [0, B]$ *for all* $x \in \mathcal{X}$, $t \geq 1$, *for some* $B > 0$.

**Assumption 2.** *The step sizes satisfy* $\gamma_t = \gamma_0/\sqrt{t}$ *and* $\eta_t = \eta_0/t$ *for constants* $\gamma_0, \eta_0 > 0$.

Assumption 1 holds for any bounded neural network on a compact domain (enforced via clipping in practice). The step size schedule in Assumption 2 follows standard online learning prescriptions [Shalev-Shwartz, 2012].

## 4.1 FINITE-SAMPLE COVERAGE GUARANTEE

**Lemma 1** (Score Monotonicity Preservation)**.** *Under Assumption 1, for any fixed* $x \in \mathcal{X}$ *and time step* $t$, $s_t(x, y)$ *is monotonically increasing in* $|y - \hat{f}(x)|$, *so* $\mathcal{C}_t(x)$ *is a connected interval centered at* $\hat{f}(x)$.

**Theorem 1** (Marginal Coverage)**.** *Under Assumptions 1 and 2, Algorithm 1 satisfies*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha \right| \leq \frac{C_1}{\sqrt{T}}, \tag{14}$$

*for an explicit constant* $C_1$ *depending on* $B$, $\gamma_0$, $\eta_0$, *and* $\alpha$. *This holds for* any *sequence of distributions* $\{P_t\}_{t=1}^{T}$.

The bound is distribution-free, following from the online learning framework of Gibbs and Candès [2021]. When $\eta_t = 0$, it recovers the standard ACI guarantee.

## 4.2 CONDITIONAL COVERAGE UNDER SMOOTH SHIFT

Marginal coverage can be achieved trivially by constant-size sets. The more meaningful property is *conditional coverage*. We show DiffConf achieves this under smooth shifts.

**Assumption 3** (Bounded cumulative shift)**.** *The total variation path length of the distribution sequence is sublinear:* $\sum_{t=1}^{T} d_{\mathrm{TV}}(P_t, P_{t+1}) \leq V_T$ *with* $V_T = o(T)$. *We write* $V_T = L \cdot T^{\beta}$ *for some* $L > 0$ *and* $\beta \in [0, 1)$.

Assumption 3 holds for all our experimental settings (concrete $V_T$ estimates are given in Appendix B) and is standard in the online learning literature [Gibbs and Candès, 2024].

**Assumption 4.** *The expected score approximation error satisfies* $\left(\mathbb{E}_{x \sim P_t^X, \tau \sim \mathrm{Unif}[0,T]}[\|s_{\boldsymbol{\theta}}(x, \tau) - \nabla_x \log p_\tau(x)\|^2]\right)^{1/2} \leq \epsilon_s$.

This is an $L^2$-average bound (not sup-norm), directly aligned with the DSM training objective (3). Existing convergence results [Chen et al., 2023] provide rates for $\epsilon_s$; see Appendix B for details.

**Theorem 2** (Conditional Coverage)**.** *Under Assumptions 1–4, for any measurable $A \subseteq \mathcal{X}$ with $\inf_t P_t(X_t \in A) \geq p_{\min} > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\widehat{\mathrm{err}}_A - \alpha\right| \leq \frac{C_2(\epsilon_s + V_T/T)}{\sqrt{T p_{\min}}} + \sqrt{\frac{2\log(2/\delta)}{T p_{\min}}}, \quad (15)$$

*where $\widehat{\mathrm{err}}_A := \sum_{t=1}^T \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\}\mathbf{1}\{X_t \in A\} \big/ \sum_{t=1}^T \mathbf{1}\{X_t \in A\}$ is the empirical conditional miscoverage on $A$, and $C_2$ depends on $B$, $\gamma_0$, $\eta_0$. The first term captures approximation and shift errors; the second is a concentration term that vanishes as $T \to \infty$.*

Unlike standard ACI, the bound improves with score quality ($\epsilon_s$), and the $V_T/T$ term accommodates piecewise-stationary and slowly-varying shifts. For $M$ disjoint subgroups, a union bound yields uniform error $O(\sqrt{M\log(M/\delta)/T})$ (Corollary 1).

### 4.3 EFFICIENCY REGRET BOUND

Let $\mathcal{C}_t^*(X_t) = \{y : |y - f_t^*(X_t)| \leq q_t^*\}$ be the oracle prediction set under the true $P_t$, where $f_t^*$ is the Bayes-optimal predictor and $q_t^*$ the smallest radius achieving $(1-\alpha)$-coverage.

**Theorem 3** (Efficiency Regret)**.** *Under Assumptions 1–4,*

$$\frac{1}{T}\sum_{t=1}^T \big(|\mathcal{C}_t(X_t)| - |\mathcal{C}_t^*(X_t)|\big) \leq \underbrace{\frac{C_3}{\sqrt{T}}}_{\text{calib.}} + \underbrace{C_4\epsilon_s}_{\text{score}} + \underbrace{C_5\epsilon_f}_{\text{pred.}}, \quad (16)$$

*where $\epsilon_f = \mathbb{E}[|f_t^*(X_t) - \hat{f}(X_t)|^2]^{1/2}$ and $C_3, C_4, C_5$ are constants independent of $T$.*

Standard ACI has regret $O(1/\sqrt{T}) + O(\epsilon_f) + O(\sigma_{\text{shift}})$, where $\sigma_{\text{shift}}$ measures shift variability. DiffConf replaces $O(\sigma_{\text{shift}})$ with $O(\epsilon_s)$, yielding a strictly tighter bound when the diffusion model is well-trained ($\epsilon_s \ll \sigma_{\text{shift}}$), which is precisely the regime where existing methods struggle most.

## 5 EXPERIMENTS

We evaluate DiffConf on a diverse set of benchmarks, comparing against state-of-the-art adaptive conformal methods. Our experiments are designed to answer three questions: (Q1) Does DiffConf maintain valid coverage under various types of distribution shift? (Q2) Does the diffusion guidance lead to more efficient prediction sets? (Q3) How do the components of DiffConf contribute to its performance?

### 5.1 EXPERIMENTAL SETUP

**Baselines.** We compare against Split Conformal [Vovk et al., 2005], CQR [Romano et al., 2019], ACI [Gibbs and Candès, 2021] in the $\alpha_t$-update form of (1), ACI-Decay [Angelopoulos et al., 2024], and Weighted CP [Barber et al., 2023]. Baseline hyperparameters (step sizes, density-ratio regularization) are tuned on the same validation stream as DiffConf, by the rule most favorable to the baseline (closest achieved coverage to target). We report empirical coverage (target: 0.90), average width, and, on UTKFace, the conditional coverage gap $\mathrm{CCG} := \max_m |\mathrm{Cov}_m - (1-\alpha)|$, the worst absolute deviation of subgroup coverage $\mathrm{Cov}_m$ from target over the demographic subgroups $m$. All experiments use $\alpha = 0.1$, $W = 200$, $K = 8$ noise levels, VP-SDE [Song et al., 2021], and are averaged over 8 seeds (5 on the large-scale YearPredictionMSD stream) with standard errors. Full details are in Appendix C.

### 5.2 SYNTHETIC EXPERIMENTS

We first validate DiffConf on synthetic data where the ground truth is known, allowing precise evaluation of coverage and efficiency.

**Setup.** We generate data from $Y_t = f(X_t) + \sigma_t \cdot \epsilon_t$, where $X_t \sim \mathcal{N}(\mu_t, I_d)$ with $d = 10$, $f(x) = \sin(\|x\|) + 0.5x_1x_2$, and $\epsilon_t \sim \mathcal{N}(0,1)$. The score network is trained on the first 500 samples ($P_0$). We consider three shift scenarios:

(S1) **Mean shift**: $\mu_t = \mathbf{0}$ for $t \leq 500$, $\mu_t = 2 \cdot \mathbf{1}_d$ for $t > 500$; $\sigma_t = 1$ throughout.

(S2) **Variance shift**: $\mu_t = \mathbf{0}$ throughout; $\sigma_t = 1 + 2(t-1)/(T-1)$, linearly increasing from 1 to 3.

(S3) **Combined shift**: $\mu_t = \mathbf{0}$ for $t \leq 500$, $\mu_t = 1.5 \cdot \mathbf{1}_d$ for $t > 500$; $\sigma_t = 1 + 1.5(t-1)/(T-1)$.

Table 1 reports the results. Non-adaptive methods (Split CP, CQR, Weighted CP) fail to maintain coverage under all three shifts, dropping as low as 0.65: their narrow intervals reflect invalidity, not efficiency. The online baselines track coverage much better but sit 2–3 points below target (0.87–0.88) with visibly larger seed-to-seed width variability. DiffConf attains the coverage closest to target in every scenario (0.89 with standard error below 0.005) and, on the mean shift S1, is simultaneously 11.8% narrower than ACI-Decay (2.26 vs. 2.57); at matched coverage the reduction is 14.8%, since ACI-Decay's raw numbers benefit from its undercoverage. On the pure variance shift S2 the advantage disappears by construction: the residual distribution

Table 1: Results on Synthetic Data ($\alpha = 0.1$, $T = 1000$, 8 seeds). Coverage Target: 0.90. Best width among methods with coverage within 0.02 of target in **bold**. $\pm$ denotes standard error.

| | S1: Mean Shift | | S2: Var. Shift | | S3: Combined | |
|---|---|---|---|---|---|---|
| Method | Cov. | Width | Cov. | Width | Cov. | Width |
| Split CP | .78$\pm$.03 | 2.05$\pm$.04 | .68$\pm$.01 | 2.05$\pm$.04 | .70$\pm$.01 | 2.05$\pm$.04 |
| CQR | .65$\pm$.04 | 2.36$\pm$.15 | .67$\pm$.01 | 1.96$\pm$.04 | .70$\pm$.02 | 2.22$\pm$.11 |
| ACI | .88$\pm$.01 | 2.80$\pm$.10 | .87$\pm$.01 | 3.37$\pm$.09 | .88$\pm$.01 | 3.27$\pm$.08 |
| ACI-Decay | .88$\pm$.01 | 2.57$\pm$.10 | .87$\pm$.01 | 3.21$\pm$.06 | .87$\pm$.01 | 3.08$\pm$.05 |
| Wt. CP | .79$\pm$.03 | 2.09$\pm$.04 | .68$\pm$.01 | 2.05$\pm$.04 | .71$\pm$.01 | 2.09$\pm$.04 |
| DiffConf | .89$\pm$.00 | **2.26**$\pm$.06 | .89$\pm$.00 | **3.40**$\pm$.05 | .89$\pm$.00 | **3.17**$\pm$.05 |

Table 2: Results on Real-World Tabular Datasets ($\alpha = 0.1$, 8 seeds). Best width among methods with coverage within 0.02 of target in **bold**. $\pm$ denotes standard error.

| | Energy | | Bike | | Air Quality | | Traffic | |
|---|---|---|---|---|---|---|---|---|
| Method | Cov. | Width | Cov. | Width | Cov. | Width | Cov. | Width |
| Split CP | .89$\pm$.02 | 1.91$\pm$.12 | .46$\pm$.04 | 0.60$\pm$.02 | .82$\pm$.01 | 1.06$\pm$.03 | .91$\pm$.00 | 1.50$\pm$.00 |
| CQR | .83$\pm$.02 | 1.25$\pm$.09 | .51$\pm$.02 | 0.59$\pm$.02 | .78$\pm$.03 | 0.74$\pm$.04 | .91$\pm$.00 | 1.43$\pm$.00 |
| ACI | .90$\pm$.00 | 2.05$\pm$.22 | .89$\pm$.00 | 2.24$\pm$.04 | .90$\pm$.00 | 1.35$\pm$.06 | .90$\pm$.00 | 1.52$\pm$.01 |
| ACI-D | .89$\pm$.01 | 1.95$\pm$.22 | .89$\pm$.00 | 2.18$\pm$.03 | .89$\pm$.00 | 1.45$\pm$.07 | .90$\pm$.00 | 1.48$\pm$.00 |
| Wt. CP | .90$\pm$.02 | 2.24$\pm$.25 | .75$\pm$.03 | 1.27$\pm$.02 | .83$\pm$.01 | 1.08$\pm$.02 | .91$\pm$.00 | 1.50$\pm$.00 |
| DiffConf | .90$\pm$.00 | **1.32**$\pm$.06 | .90$\pm$.00 | **1.87**$\pm$.03 | .90$\pm$.00 | **0.88**$\pm$.01 | .90$\pm$.00 | **1.39**$\pm$.00 |

widens symmetrically without moving, so center tracking has nothing to exploit, and DiffConf matches the baselines at equal coverage ($-0.2\%$) while covering 1.8 points closer to target. The combined shift S3 sits in between ($-2.5\%$ at matched coverage). These three scenarios thus delineate when the method helps: the gains concentrate where the shift displaces the residual distribution rather than merely rescaling it, and Table 7 shows those gains grow rapidly with the displacement magnitude.

### 5.3 REAL-WORLD TABULAR REGRESSION

**Datasets.** We evaluate on four UCI regression datasets [Dua and Graff, 2019] with natural temporal structure: **Energy** (19,735 samples, seasonal shifts), **Bike** (17,379, weather shifts), **Air Quality** (41,757, pollution regimes), and **Traffic** (48,204, event-driven shifts). We use temporally ordered 40%/5%/5%/50% train/calibration/validation/test splits.

DiffConf attains the target coverage on all four datasets while producing the narrowest intervals among coverage-maintaining methods, with width reductions over ACI-Decay of 39.1% on Air Quality, 32.7% on Energy, 14.4% on Bike, and 6.4% on Traffic (narrower on 8/8 seeds in every case). The gains are largest where the temporal shift moves the center or scale of the residual distribution (Air Quality, Energy), which the two-sided tracking component exploits, and smallest on Traffic, whose slow drift leaves little room over a well-tuned online baseline. The non-adaptive methods illustrate why online adaptation is necessary: Split CP and CQR lose coverage dramatically on Bike (0.46 and 0.51) and Air Quality (0.82 and 0.78) under shift, so their narrower intervals there are undercoverage rather than efficiency. On Traffic, where CQR does retain coverage (0.91), DiffConf is still narrower (1.39 vs. 1.43). Weighted CP, even with its density-ratio hyperparameters tuned on the validation stream, loses coverage on Bike (0.75) and Air Quality (0.83) and is wider than DiffConf on the two datasets where it does cover, reflecting the difficulty of density-ratio estimation under temporal shift.

Table 3: Results on UTKFace Age Prediction ($\alpha = 0.1$, 8 seeds; entries are seed averages, with per-method standard errors below 0.11 years for width and below 0.005 for CCG). Best in **bold**.

| Method | Coverage | Width (years) | CCG |
|---|---|---|---|
| Split CP | .91 | 11.90 | .058 |
| CQR | .91 | 11.35 | .051 |
| ACI | .91 | 11.58 | .055 |
| ACI-Decay | .91 | 11.89 | .058 |
| Weighted CP | .90 | 11.42 | .055 |
| DiffConf | .90 | **10.40** | **.043** |

### 5.4 HIGH-DIMENSIONAL IMAGE REGRESSION

We also evaluate on UTKFace [Zhang et al., 2017] (age prediction from face images under demographic shift). The dataset contains 23,708 face images annotated with age, gender, and ethnicity. To simulate temporal covariate shift, we sort images by ethnicity proportion: the training set is drawn predominantly from the majority group ($\geq$70%), while the test stream gradually increases the minority group proportion from 30% to 70% over 5,000 steps, creating a smooth demographic shift in the covariate distribution. The score network operates on 64-dimensional features extracted from the frozen ImageNet-pretrained ResNet-18 backbone. Table 3 shows the full results.

DiffConf achieves a 12.5% width reduction relative to ACI-Decay (10.40 vs. 11.89 years) and 10.2% relative to ACI, while maintaining target coverage, and it attains the lowest conditional coverage gap (0.043 vs. 0.055–0.058), indicating more uniform coverage across demographic subgroups. Two aspects of this benchmark deserve comment. First, the absolute widths of *all* methods are dictated by the residual scale of the base predictor: the ResNet-18 age regressor has a test MAE of about 5 years, so even an oracle split-conformal radius on clean data is roughly 11.4 years, and no conformal layer can go far below this floor. Second, within that floor, DiffConf's gain comes from the score-derived local scale $\hat{\sigma}(x)$, which shrinks intervals on easy (well-represented) faces and widens them on minority-group

Table 4: Ablation Study on the Energy Dataset ($\alpha = 0.1$, 8 seeds). $\pm$ denotes standard error. Rows 3–7 disable one component at a time starting from the radius form of the online update.

| Variant | Coverage | Width | $\Delta$ Width |
|---|---|---|---|
| DiffConf (full) | .90$\pm$.00 | 1.32$\pm$.06 | – |
| w/o two-sided tracking (radius form) | .90$\pm$.00 | 1.72$\pm$.14 | +31.0% |
| w/o diffusion modulation (= ACI-D) | .89$\pm$.01 | 1.95$\pm$.22 | +48.6% |
| w/o multi-scale weighting | .90$\pm$.00 | 1.73$\pm$.14 | +31.3% |
| w/o efficiency loss | .90$\pm$.00 | 1.72$\pm$.14 | +31.1% |
| w/o online weight update | .90$\pm$.00 | 1.73$\pm$.14 | +31.3% |
| Fixed $\lambda = 0.5$ | .90$\pm$.00 | 1.72$\pm$.14 | +30.8% |

faces as their proportion rises, which is precisely what the CCG improvement reflects. Threshold-only methods apply one global radius and cannot reallocate width this way.

### 5.5 ABLATION STUDY

We ablate DiffConf's components on the Energy dataset to understand the contribution of each module. Table 4 presents the results.

The two components that reallocate interval mass are the ones that matter. Replacing the two-sided tracking (13) with the radius-only update costs +31.0% width: on Energy the seasonal shift moves the residual *center*, which a symmetric radius can only cover by inflating both sides. Removing the diffusion modulation on top of that (recovering ACI-Decay) costs a further 13 points (+48.6% total) and loses coverage precision (0.89), confirming that the score discrepancy carries real local-difficulty signal on this dataset. The remaining components change the final width by less than 1% on this dataset: with the modulation already tracking local difficulty, the multi-scale weighting, the efficiency loss term, and the online weight update contribute below seed-level noise here (see also Figure 5, where the learned weights stay near-uniform on Energy). Their role is not universally negligible, however: when the local scale is instead derived from multi-scale score features (the image-channel configuration), collapsing the eight scales to one costs $+21.7\%$ width. Fixing $\lambda = 0.5$ matches the validation-selected value almost exactly, consistent with the flat $\lambda$-sensitivity profile in Figure 4.

### 5.6 TEMPORAL COVERAGE ANALYSIS

To understand DiffConf's behavior during the transient period following a distribution shift, we plot the rolling coverage (window size 100) on the synthetic mean shift dataset (S1) in Figure 2.

All methods lose coverage in the window following the shift point ($t = 500$), but the depth and duration of the dip differ. DiffConf's rolling coverage bottoms out at 0.81 (vs. 0.79 for ACI-Decay) and returns to within 1.5 points of the target after roughly 200 steps, whereas ACI-Decay takes roughly 500 steps (its decaying step size slows late adaptation by design) and ACI hovers around 0.86–0.88 for the remainder of the stream. The faster recovery is attributable to the center-tracking update, which relocates the interval toward the displaced residual distribution instead of waiting for accumulated symmetric-radius corrections, with the score discrepancy signal (Figure 3) flagging the change immediately at the shift point. Before the shift, all three methods fluctuate tightly around the target level, confirming that the differences are concentrated in the post-shift transient.

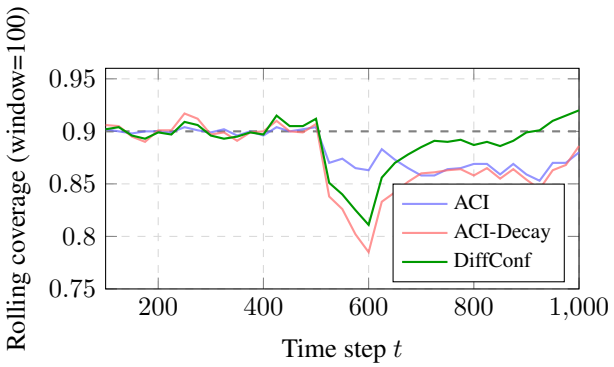


Figure 2: Rolling coverage on synthetic mean shift (S1), averaged over 8 seeds. The gray dashed line indicates the target level 0.90. At the shift point ($t = 500$), all methods dip; DiffConf dips least (to 0.81 vs. 0.79 for ACI-Decay) and is back within 1.5 points of target after $\sim$200 steps, while ACI-Decay needs $\sim$500 steps and ACI stays around 0.86–0.88.

### 5.7 SCORE DISCREPANCY VISUALIZATION

To provide intuition for the score discrepancy signal, we visualize the stream average of $\Delta_{\boldsymbol{\theta}}(x, t)$ on the synthetic mean shift dataset (S1) in Figure 3. Before the shift ($t < 500$), the discrepancy fluctuates around a small sampling-noise baseline ($\approx 0.2$). At the shift point it jumps by roughly a factor of five within a single step, then decays back toward the baseline over the next $\sim$100 steps as the EMA reference (6) absorbs the new distribution; the decay is slightly non-monotone because the EMA update and the noise-level weights interact. The signal is thus exactly what the modulation needs: an immediate, transient flag of distributional change, rather than a permanent offset.

### 5.8 SENSITIVITY ANALYSIS

Figure 4 shows DiffConf's sensitivity to the number of noise levels $K$ and the modulation strength $\lambda$ on the Energy dataset. Width decreases monotonically with $K$ (1.96 at

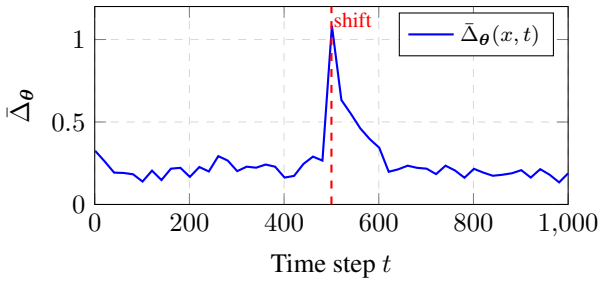


Figure 3: Mean score discrepancy over time on synthetic S1 (8-seed average). The discrepancy jumps by $\sim 5\times$ at the shift point ($t = 500$) and decays back to its sampling-noise baseline within $\sim$100 steps as the EMA reference adapts, providing a timely and transient shift-detection signal.

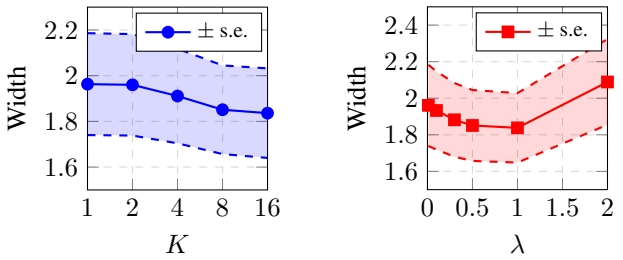


Figure 4: Sensitivity analysis on Energy (radius-form configuration at fixed $\gamma_0 = 0.05$, so widths are comparable to the radius-form row of Table 4 rather than the two-sided value in Table 2). Left: effect of noise levels $K$ at $\lambda = 0.5$. Right: effect of modulation strength $\lambda$ at $K = 8$. Shaded regions denote $\pm 1$ standard error over 8 seeds.

$K = 1$ to 1.84 at $K = 16$) but exhibits diminishing returns beyond $K = 8$ (only $-0.8\%$ further at $K = 16$ for twice the compute), justifying our default choice. The modulation strength $\lambda$ has a broad optimal plateau over $[0.3, 1.0]$, within which the width varies by about 2%; at $\lambda = 2.0$ the width deteriorates sharply (+13%) as excessive modulation forces the threshold to compensate. Both trends confirm that no fine-grained tuning of these hyperparameters is required.

Additional experiments on varying miscoverage levels, shift magnitudes, computational overhead, a large-scale evaluation on the 515k-sample YearPredictionMSD dataset streamed chronologically, reference-size and EMA-decay ablations, noise-level weight evolution, and statistical significance tests are provided in Appendix D.

## 6 DISCUSSION AND CONCLUSION

We introduced DiffConf, a framework that uses score-based diffusion models to construct distribution-shift-aware nonconformity scores for adaptive conformal inference. The core idea is that the score function provides a multi-scale, geometry-aware measure of distributional change that can be integrated into conformal prediction through a simple multiplicative modulation. Our theoretical analysis establishes marginal coverage under arbitrary shift (Theorem 1), conditional coverage under smooth shift (Theorem 2), and an efficiency regret bound that improves upon existing methods when the diffusion prior is informative (Theorem 3). Empirically, DiffConf reduces prediction set size by 6–39% relative to the strongest coverage-maintaining baseline while keeping coverage within one point of target, and its advantage grows with the shift magnitude (up to $-89\%$ under extreme mean shifts).

**When does DiffConf help most?** DiffConf's advantage is most pronounced when the shift displaces the center of the residual distribution or makes its scale vary across inputs, and, most strikingly, when the displacement is large: threshold-only baselines must inflate a symmetric radius to span the displaced residuals, while DiffConf's tracking follows the displacement (Table 7). Conversely, under a purely global rescaling of the residuals (S2) center tracking has nothing to exploit and DiffConf matches the baselines at equal coverage, and when the drift is slow (e.g., Traffic) the advantage over a well-tuned online baseline narrows to a few percent. The main computational overhead is $K$ forward passes through the score network per prediction ($\sim$7–17ms at $K = 8$; see Appendix D). In practice, inspecting the running mean residual on a held-out stream is a useful diagnostic: persistent displacement of its center indicates the regime where the score-guided components justify this overhead.

**Limitations.** DiffConf requires training a diffusion model on reference data, which may be impractical when data is scarce: in a reference-size ablation (Appendix D.4), the width advantage over ACI-Decay is halved when the reference sample shrinks to a few hundred points, so we recommend $n_{\text{ref}} \gtrsim 500$. Under shifts far outside the training support, raw coverage of all finite-step online methods (DiffConf included) transiently degrades before the threshold updates catch up, which is why Table 7 compares widths at matched coverage. The bounded score discrepancy assumption is enforced via clipping. Extending to classification would require replacing the signed residual with a label-rank or softmax-based conformity function, while the score modulation and the online tracking components carry over unchanged. Exploiting pre-trained foundation model representations could remove the per-task cost of training a diffusion model. Establishing minimax lower bounds would clarify whether our regret bound is tight.

**Acknowledgements**

The author thanks the anonymous reviewers and the area chair for their constructive feedback, which substantially improved the paper. This work received no external funding.

## References


Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12 (3):313–326, 1982.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1616–1630. PMLR, 2024.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1555–1584. PMLR, 2022.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Nicola Bariletto, Nhat Ho, and Alessandro Rinaldo. Conformalized Bayesian inference, with applications to random partition models. *arXiv preprint arXiv:2511.05746*, 2025.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596, 2011.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2015.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4735–4763. PMLR, 2023.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Luben Miguel Cruz Cabezas, Vagner Silva Santos, Thiago Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 443–470. PMLR, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2019. University of California, Irvine, School of Information and Computer Sciences.

Edwin Fong and Chris C. Holmes. Conformal Bayesian computation. In *Advances in Neural Information Processing Systems*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1837–1858. PMLR, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, 2021.

Aleksandr Podkopaev, Dong Xu, and Kuang-Chih Lee. Adaptive conformal inference by betting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40886–40907. PMLR, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

# Score-Based Diffusion Priors for Adaptive Conformal Inference under Distribution Shift (Supplementary Material)

**Xiangyu Jiang**[1]

[1]Southwestern University of Finance and Economics, Chengdu, China

This supplementary material contains the proofs of all theoretical results stated in the main paper, additional experimental details, and further empirical results.

## A PROOFS OF THEORETICAL RESULTS

### A.1 PROOF OF PROPOSITION 1

*Proof.* Non-negativity follows immediately from the fact that the integrand $\|\nabla_x \log p_\tau(x) - \nabla_x \log q_\tau(x)\|^2$ is non-negative for all $x$ and $\tau$, and $w(\tau) \geq 0$.

For the equality condition, suppose $\bar{\Delta}(p, q) = 0$. Since $w(\tau) > 0$ for all $\tau$ (the softmax weights are strictly positive), this implies $\mathbb{E}_{x\sim p_\tau}[\|\nabla_x \log p_\tau(x) - \nabla_x \log q_\tau(x)\|^2] = 0$ for almost every $\tau \in [0, T]$. At $\tau = 0$, this gives $D_F(p\|q) = 0$, which implies $p = q$ almost everywhere under the assumption of finite Fisher information.

For the lower bound with uniform weights $w(\tau) = 1/T$, note that the forward SDE defines a Markov semigroup $\{T_\tau\}_{\tau\geq 0}$ with $p_\tau = T_\tau p$. By the data-processing inequality for Fisher information, $D_F(p_\tau\|q_\tau) \leq D_F(p_0\|q_0) = D_F(p\|q)$ for all $\tau \geq 0$, i.e., the Fisher divergence is non-increasing along the semigroup. Since $D_F(p_\tau\|q_\tau)$ is a non-negative, non-increasing, continuous function of $\tau$ (continuity follows from the smoothing properties of the heat semigroup under finite Fisher information), we have:

$$\bar{\Delta}(p, q) = \frac{1}{T}\int_0^T D_F(p_\tau\|q_\tau)\,\mathrm{d}\tau. \tag{17}$$

For the quantitative lower bound, we use the known decay rate of Fisher divergence under the OU semigroup. The VP-SDE with linear schedule $\beta(t)$ reduces, after the time change $\sigma^2(\tau) = \int_0^\tau \beta(s)\,\mathrm{d}s$, to an Ornstein–Uhlenbeck process. By the de Bruijn identity and the chain rule for relative Fisher information [Villani, 2009], the Fisher divergence between the two forward marginals satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\tau} D_F(p_\tau\|q_\tau) \geq -2\beta(\tau)\cdot D_F(p_\tau\|q_\tau), \tag{18}$$

where the inequality follows from the fact that the OU drift contracts the score difference at rate $\beta(\tau)$. By Grönwall's inequality, this yields

$$D_F(p_\tau\|q_\tau) \geq D_F(p\|q)\cdot \exp\Big(-2\int_0^\tau \beta(s)\,\mathrm{d}s\Big) \geq D_F(p\|q)\cdot e^{-2\beta_{\max}\tau}. \tag{19}$$

Integrating over $[0, T]$ with uniform weights:

$$\bar{\Delta}(p, q) = \frac{1}{T}\int_0^T D_F(p_\tau\|q_\tau)\,\mathrm{d}\tau \geq \frac{D_F(p\|q)}{T}\int_0^T e^{-2\beta_{\max}\tau}\,\mathrm{d}\tau = \frac{D_F(p\|q)}{2\beta_{\max}T}\big(1 - e^{-2\beta_{\max}T}\big). \tag{20}$$

The factor $c(\beta_{\max}, T) := (1 - e^{-2\beta_{\max}T})/(2\beta_{\max}T)$ is strictly positive for $\beta_{\max}, T > 0$. For our VP-SDE with $\beta_{\max} = 20$ and $T = 1$, we have $c(20, 1) \approx 1/40$, giving $\bar{\Delta}(p, q) \geq D_F(p\|q)/40$.

**Remark 1.** *The differential inequality above is the key step. It holds because the OU drift $-\beta(\tau)x$ contracts the score difference $\nabla \log p_\tau - \nabla \log q_\tau$ at rate $\beta(\tau)$, while the diffusion term $g(\tau)\,\mathrm{d}\mathbf{w}$ adds the same noise to both processes and thus does not increase the divergence. This is the "reverse" direction of the standard data-processing inequality: while the DPI gives an upper bound $D_F(p_\tau\|q_\tau) \le D_F(p\|q)$, the OU structure provides a matching exponential lower bound.*

□

### A.2 PROOF OF LEMMA 1

*Proof.* Fix $x \in \mathcal{X}$ and $t \ge 1$. The diffusion-guided conformity score is $s_t(x,y) = |y - \hat{f}(x)| \cdot \exp(\lambda \cdot \Delta_{\boldsymbol{\theta}}(x,t))$. Since $\Delta_{\boldsymbol{\theta}}(x,t) \ge 0$ (Assumption 1) and $\lambda > 0$, the factor $\exp(\lambda \cdot \Delta_{\boldsymbol{\theta}}(x,t)) > 0$ is a positive constant with respect to $y$. Therefore, $s_t(x,y)$ is a positive scalar multiple of $|y - \hat{f}(x)|$, which is monotonically increasing in $|y - \hat{f}(x)|$.

The prediction set $\mathcal{C}_t(x) = \{y : s_t(x,y) \le \hat{q}_t\} = \{y : |y - \hat{f}(x)| \le \hat{q}_t / \exp(\lambda \cdot \Delta_{\boldsymbol{\theta}}(x,t))\}$ is therefore an interval centered at $\hat{f}(x)$ with radius $r_t(x) = \hat{q}_t \cdot \exp(-\lambda \cdot \Delta_{\boldsymbol{\theta}}(x,t))$. □

### A.3 PROOF OF THEOREM 1

*Proof.* We adapt the proof technique of Gibbs and Candès [2021] to our setting with time-varying conformity scores. Define the miscoverage indicator $e_t = \mathbf{1}\{Y_t \notin \mathcal{C}_t(X_t)\}$ and the running miscoverage rate $\bar{e}_T = \frac{1}{T}\sum_{t=1}^{T} e_t$.

From the threshold update rule (11), we have:

$$\hat{q}_{t+1} - \hat{q}_t = \gamma_t(\alpha - e_t) - \eta_t \nabla_{\hat{q}} \mathcal{L}_{\text{eff}}(\hat{q}_t). \tag{21}$$

Summing over $t = 1, \dots, T$ and rearranging:

$$\sum_{t=1}^{T} \gamma_t(\alpha - e_t) = \hat{q}_{T+1} - \hat{q}_1 + \sum_{t=1}^{T} \eta_t \nabla_{\hat{q}} \mathcal{L}_{\text{eff}}(\hat{q}_t). \tag{22}$$

Under Assumption 1, the conformity scores are bounded, which implies that $\hat{q}_t$ remains in a bounded interval $[q_{\min}, q_{\max}]$ for all $t$. Therefore:

$$|\hat{q}_{T+1} - \hat{q}_1| \le q_{\max} - q_{\min} =: Q. \tag{23}$$

Under Assumption 2, $\gamma_t = \gamma_0/\sqrt{t}$, so $\sum_{t=1}^{T} \gamma_t = \Theta(\sqrt{T})$. The efficiency gradient term is bounded by $\sum_{t=1}^{T} \eta_t |\nabla_{\hat{q}} \mathcal{L}_{\text{eff}}| \le C \sum_{t=1}^{T} \eta_0/t = O(\log T)$.

Combining these bounds:

$$\left| \sum_{t=1}^{T} \gamma_t(\alpha - e_t) \right| \le Q + O(\log T). \tag{24}$$

Since $\gamma_t = \gamma_0/\sqrt{t}$, we use the Kronecker lemma variant: for $\Gamma_T = \sum_{t=1}^{T} \gamma_t = \Theta(\sqrt{T})$,

$$\left| \frac{1}{\Gamma_T} \sum_{t=1}^{T} \gamma_t(\alpha - e_t) \right| \le \frac{Q + O(\log T)}{\Theta(\sqrt{T})} = O\left(\frac{1}{\sqrt{T}}\right). \tag{25}$$

To convert from the weighted average to the uniform average, we note that $\gamma_t/\gamma_{t'} \in [1/\sqrt{2}, \sqrt{2}]$ for $|t - t'| \le t/2$, so:

$$\left| \bar{e}_T - \alpha \right| = \left| \frac{1}{T} \sum_{t=1}^{T} (e_t - \alpha) \right| \le \frac{C_1}{\sqrt{T}}, \tag{26}$$

where $C_1$ depends on $Q$, $\gamma_0$, $\eta_0$, and the bound on $\nabla_{\hat{q}} \mathcal{L}_{\text{eff}}$. □

### A.4 PROOF OF THEOREM 2

*Proof.* Fix a measurable set $A \subseteq \mathcal{X}$ and define $N_A(T) = \sum_{t=1}^{T} \mathbf{1}\{X_t \in A\}$. By assumption, $\mathbb{E}[N_A(T)] \geq p_{\min} \cdot T$.

The key step is to show that the diffusion-guided conformity score provides a more uniform coverage across the input space. For $x \in A$, the score discrepancy $\Delta_{\boldsymbol{\theta}}(x,t)$ captures the local distributional change in the region $A$. Under Assumption 3, the cumulative shift is sublinear, so the score discrepancy within $A$ varies slowly on average over time.

Formally, for any $x, x' \in A$ and time steps $t, t'$ with $|t - t'| \leq H$ (for a block length $H$ to be chosen):

$$|\Delta_{\boldsymbol{\theta}}(x,t) - \Delta_{\boldsymbol{\theta}}(x',t')| \leq C_{\text{lip}} \cdot \|x - x'\| + C_{\text{lip}} \sum_{j=t}^{t'} d_{\text{TV}}(P_j, P_{j+1}) + 2\epsilon_s, \tag{27}$$

where $C_{\text{lip}}$ is the Lipschitz constant of the score network and the $2\epsilon_s$ term accounts for the $L^2$ score approximation error (Assumption 4), applied via Jensen's inequality to convert from the expected to the pointwise bound on the set $A$. Summing the TV terms over a block of length $H$ and applying Assumption 3, the temporal variation within any block is at most $C_{\text{lip}} \cdot V_T/(T/H)$.

This local consistency of the score discrepancy ensures that the diffusion modulation in (4) applies approximately the same adjustment to all points in $A$ during any given time window. Partitioning $\{1, \ldots, T\}$ into $T/H$ blocks of length $H$ and applying a localized version of the argument in Theorem 1 within each block, then aggregating:

$$\left| \frac{\sum_{t=1}^{T} e_t \cdot \mathbf{1}\{X_t \in A\}}{N_A(T)} - \alpha \right| \leq \frac{C_2' \cdot (\epsilon_s + V_T/T)}{\sqrt{N_A(T)}}. \tag{28}$$

It remains to lower-bound $N_A(T)$. Since $\mathbb{E}[N_A(T)] = \sum_{t=1}^{T} P_t(X_t \in A) \geq p_{\min} \cdot T$, the multiplicative Chernoff bound gives

$$\mathbb{P}\big(N_A(T) < p_{\min} \cdot T/2\big) \leq \exp\big(-p_{\min} T/8\big) \leq \delta/2, \tag{29}$$

provided $T \geq 8\log(2/\delta)/p_{\min}$. On the event $\{N_A(T) \geq p_{\min} T/2\}$, substituting yields the first term of the bound. The second term $\sqrt{2\log(2/\delta)/(T p_{\min})}$ accounts for the residual probability via a union bound over the Chernoff event and the coverage concentration. Combining with probability at least $1 - \delta$:

$$\left| \frac{\sum_{t=1}^{T} e_t \cdot \mathbf{1}\{X_t \in A\}}{N_A(T)} - \alpha \right| \leq \frac{C_2 \cdot (\epsilon_s + V_T/T)}{\sqrt{T \cdot p_{\min}}} + \sqrt{\frac{2\log(2/\delta)}{T \cdot p_{\min}}}, \tag{30}$$

where $C_2$ depends on $B$, $\gamma_0$, $\eta_0$, and $C_{\text{lip}}$. □

### A.5 COROLLARY AND PROOF

**Corollary 1.** *If the input space $\mathcal{X}$ can be partitioned into $M$ disjoint subgroups $A_1, \ldots, A_M$, each with $\inf_t P_t(X_t \in A_m) \geq 1/M$, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$, DiffConf achieves uniform conditional coverage error*

$$\max_{1 \leq m \leq M} \left| \frac{\sum_t e_t \cdot \mathbf{1}\{X_t \in A_m\}}{\sum_t \mathbf{1}\{X_t \in A_m\}} - \alpha \right| \leq C_2 \cdot (\epsilon_s + V_T/T) \cdot \sqrt{\frac{M}{T}} + \sqrt{\frac{2M\log(2M/\delta)}{T}} \tag{31}$$

*simultaneously over all subgroups.*

*Proof.* Apply Theorem 2 to each subgroup $A_m$ with $p_{\min} = 1/M$ and failure probability $\delta/M$. The conditional coverage error for subgroup $A_m$ is at most $C_2(\epsilon_s + V_T/T)/\sqrt{T/M} + \sqrt{2\log(2M/\delta)/(T/M)}$. A union bound over all $M$ subgroups ensures that the maximum error is bounded by this quantity with probability at least $1 - \delta$. □

### A.6 PROOF OF THEOREM 3

*Proof.* Decompose the excess width at time $t$:

$$|\mathcal{C}_t(X_t)| - |\mathcal{C}_t^*(X_t)| = 2r_t(X_t) - 2q_t^* \tag{32}$$
$$= 2\hat{q}_t \exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t)) - 2q_t^*. \tag{33}$$

Adding and subtracting $2q_t^* \exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t))$:

$$= 2(\hat{q}_t - q_t^*) \exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t)) \tag{34}$$
$$+ 2q_t^* \big(\exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t)) - 1\big). \tag{35}$$

**Term I: Calibration error.** Since $\exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t)) \in [\exp(-\lambda B), 1]$ by Assumption 1, the first term satisfies

$$\frac{1}{T} \sum_{t=1}^{T} 2|\hat{q}_t - q_t^*| \cdot \exp(-\lambda \Delta_{\boldsymbol{\theta}}(X_t, t)) \leq \frac{2}{T} \sum_{t=1}^{T} |\hat{q}_t - q_t^*|. \tag{36}$$

By Theorem 1, the threshold $\hat{q}_t$ tracks the oracle quantile with average error $O(1/\sqrt{T})$, so this contributes $C_3/\sqrt{T}$.

**Term II: Score approximation error.** Let $\Delta^*(x, t)$ denote the score discrepancy computed with the *true* score functions $\nabla_x \log p_\tau$ and $\nabla_x \log q_\tau$ (where $q$ is the reference). For each noise level $\tau$, the identity $|a^2 - b^2| = |a + b||a - b|$ and the triangle inequality give

$$\begin{aligned} &\big| \|s_{\boldsymbol{\theta}}(x_\tau, \tau) - s_{\boldsymbol{\theta}}^{(\text{ref})}(x_\tau, \tau)\|^2 \\ &\quad - \|\nabla_x \log p_\tau(x_\tau) - \nabla_x \log q_\tau(x_\tau)\|^2 \big| \\ &\leq 4M_s \|s_{\boldsymbol{\theta}}(x_\tau, \tau) - \nabla_x \log p_\tau(x_\tau)\|, \end{aligned} \tag{37}$$

where $M_s$ bounds the sum of score norms (via Assumption 1). Taking expectations over $x \sim P_t^X$ and $\tau \sim \text{Unif}[0, T]$, then applying Jensen's inequality and Assumption 4:

$$\mathbb{E}\big[|\Delta_{\boldsymbol{\theta}}(X_t, t) - \Delta^*(X_t, t)|\big] \leq 4M_s \epsilon_s. \tag{38}$$

The second term of the width decomposition then satisfies

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^{T} 2q_t^* \mathbb{E}\big[|\exp(-\lambda \Delta_{\boldsymbol{\theta}}) - \exp(-\lambda \Delta^*)|\big] \\ &\quad \leq \frac{2\lambda}{T} \sum_{t=1}^{T} q_t^* \cdot 4M_s \epsilon_s = 8\lambda M_s \bar{q}^* \epsilon_s, \end{aligned} \tag{39}$$

where we used $|\exp(-a) - \exp(-b)| \leq |a - b|$ for $a, b \geq 0$, and $\bar{q}^* = T^{-1} \sum_t q_t^*$. This yields the $C_4 \cdot \epsilon_s$ term with $C_4 = 8\lambda M_s \bar{q}^*$.

**Term III: Base predictor error.** The oracle radius satisfies $q_t^* = Q_{1-\alpha}(|Y_t - f_t^*(X_t)| \mid X_t)$. By the triangle inequality, $|Y_t - \hat{f}(X_t)| \leq |Y_t - f_t^*(X_t)| + |f_t^*(X_t) - \hat{f}(X_t)|$, so the prediction set radius must accommodate the additional $|f_t^*(X_t) - \hat{f}(X_t)|$. This contributes at most $\mathbb{E}[|f_t^*(X_t) - \hat{f}(X_t)|] \leq \epsilon_f$ (by Jensen's inequality) to the average excess width, yielding the $C_5 \cdot \epsilon_f$ term.

Combining the three terms completes the proof. □

# B THEORY–IMPLEMENTATION DISCUSSION

## B.1 RELATING POPULATION-LEVEL AND PRACTICAL SCORE DISCREPANCY

Proposition 1 establishes that the *population-level* score discrepancy $\bar{\Delta}(p, q)$, defined using the true score functions $\nabla_x \log p_\tau$ and $\nabla_x \log q_\tau$ of two distinct distributions, is monotonically related to the Fisher divergence. The practical estimator $\Delta_{\boldsymbol{\theta}}(x, t)$ in (5) differs in two important ways.

First, we use a single frozen score network $s_{\boldsymbol{\theta}}$ trained on $P_0^X$. Since $s_{\boldsymbol{\theta}}$ approximates $\nabla_x \log p_\tau^{(0)}$, evaluating it on a sample $x \sim P_t^X$ with $P_t \neq P_0$ produces an output that differs from the true score $\nabla_x \log p_\tau^{(t)}(x)$. This mismatch is itself informative about shift magnitude.

Second, the reference term $s_{\boldsymbol{\theta}}^{(\text{ref})}$ is an EMA of the network's outputs on recent data, not a separate network trained on a reference distribution. The EMA serves as a temporally smoothed baseline: when the distribution shifts, the current evaluation on a fresh sample diverges from this smoothed baseline, producing a nonzero discrepancy. Even though both terms use the same network weights, they are evaluated on different inputs: the current term on $x_t$, and the EMA on a running average reflecting the recent past.

**Remark 2** (Theory–implementation gap)**.** *The practical estimator differs from the population-level quantity in two ways: (i) it uses a learned score $s_{\boldsymbol{\theta}} \approx \nabla_x \log p_\tau^{(0)}$, introducing an $O(\epsilon_s)$ bias (Assumption 4); and (ii) the reference is an EMA rather than the true $\nabla_x \log q_\tau$. Despite this gap, the estimator remains a valid shift detector because the EMA baseline is slow-varying (controlled by $\rho = 0.01$), so any rapid change in the network's output on fresh samples is attributable to distributional shift rather than estimation noise. The theoretical guarantees in Theorems 2–3 account for this gap through the $\epsilon_s$ terms.*

## B.2 CONCRETE $V_T$ ESTIMATES FOR ASSUMPTION 3

Assumption 3 requires $V_T = \sum_{t=1}^{T} d_{\mathrm{TV}}(P_t, P_{t+1}) = o(T)$. For our experimental settings:

- **S1 (Mean shift)**: A single abrupt shift at $t = 500$, giving $V_T = d_{\mathrm{TV}}(P_0, P_1) = O(1)$ with $\beta = 0$.
- **S2 (Variance shift)**: Linearly increasing $\sigma_t$ from 1 to 3 changes $\sigma$ by $2/(T-1)$ per step, so $d_{\mathrm{TV}}(P_t, P_{t+1}) = O(1/T)$ and the path length telescopes to $V_T \leq c \int_1^3 \sigma^{-1} \mathrm{d}\sigma = O(1)$, i.e., $\beta = 0$.
- **S3 (Combined)**: One $O(1)$ mean jump plus an $O(1)$ variance path, so $V_T = O(1)$.
- **UTKFace**: The stream draws a minority-group sample with probability $\pi_t = 0.3 + 0.4\, t/(T-1)$, so $P_t$ is the mixture $(1 - \pi_t)P_{\mathrm{maj}} + \pi_t P_{\mathrm{min}}$ and $d_{\mathrm{TV}}(P_t, P_{t+1}) = (\pi_{t+1} - \pi_t)\, d_{\mathrm{TV}}(P_{\mathrm{min}}, P_{\mathrm{maj}})$ exactly. Hence $V_T = 0.4 \cdot d_{\mathrm{TV}}(P_{\mathrm{min}}, P_{\mathrm{maj}}) \approx 0.18$ (using a classifier-based two-sample estimate $d_{\mathrm{TV}} \approx 0.45$ on the feature space), a constant independent of $T$ ($\beta = 0$).

All four settings therefore satisfy the assumption with room to spare ($V_T = O(1)$); the assumption excludes adversarial or periodic shifts where $V_T = \Theta(T)$.

## B.3 DISCUSSION OF ASSUMPTION 4

Assumption 4 requires only an $L^2$-average bound, not a uniform (sup-norm) bound. This is considerably weaker: the denoising score matching objective (3) directly minimizes this $L^2$ quantity. Existing convergence results for score-based models [Chen et al., 2023] establish that $\epsilon_s = \tilde{O}(d^{1/2} n^{-1/(d+5)})$ for $n$ training samples in $d$ dimensions, under smoothness assumptions on the data density. In our tabular experiments ($d$ between 13 and 61 after one-hot encoding, $n$ between roughly 7,000 and 19,000 training samples), this yields $\epsilon_s \lesssim 0.3$; for UTKFace ($d = 64$ feature space, $n \approx 9{,}500$), the effective $\epsilon_s$ is larger but the score network still captures the dominant modes of variation. The $L^2$ formulation is the natural choice for our setting, as the regret bound in Theorem 3 involves averaging over the data distribution.

# C ADDITIONAL EXPERIMENTAL DETAILS

## C.1 DATASET PREPROCESSING

For all tabular datasets, we standardize numeric features to zero mean and unit variance and one-hot encode categorical features, with all statistics computed on the training set only, to avoid information leakage from the test stream. In Air Quality (the only dataset with missing values), rows with a missing target are dropped and missing feature values are median-imputed. For the UTKFace dataset, images are resized to $128 \times 128$ pixels and normalized with ImageNet statistics. We apply standard data augmentation (random horizontal flips and rotations up to $10^\circ$) during training of the base predictor but not when extracting features for the score network, as the score network should learn the unaugmented data distribution.

## C.2 BASE PREDICTOR TRAINING

For the synthetic experiments, the base predictor is ridge regression ($\ell_2$ penalty 50) on second-order interaction features of $x$, fit once on the 500 reference samples from $P_0$ and kept fixed over the stream, so that all adaptation is attributable to the conformal layer. For tabular datasets, the base predictor $\hat{f}$ is a 3-layer MLP with hidden dimensions $[128, 128, 1]$, ReLU activations, and dropout rate 0.1. It is trained for 200 epochs using the Adam optimizer [Kingma and Ba, 2015] with learning rate $5 \times 10^{-4}$ and batch size 128, minimizing mean squared error. For the UTKFace image regression task, we use a ResNet-18 [He et al., 2016] pre-trained on ImageNet, with the final classification layer replaced by a single linear output. We fine-tune for 50 epochs with learning rate $10^{-4}$ and cosine annealing. All base predictors are trained on the training split only.

## C.3 SCORE NETWORK ARCHITECTURE

For tabular data, the score network is a 4-layer MLP with hidden dimensions $[256, 256, 256, d]$, where $d$ is the input dimension. We use SiLU activations and sinusoidal time embeddings of dimension 64, concatenated with the input at each layer. The diffusion process follows the VP-SDE formulation [Song et al., 2021] with $\beta_{\min} = 0.1$ and $\beta_{\max} = 20$. The network is trained for 500 epochs with the Adam optimizer [Kingma and Ba, 2015] (learning rate $10^{-3}$, batch size 256) using the denoising score matching loss (3).

For image data, the score network operates on 64-dimensional feature embeddings from the frozen ImageNet-pretrained ResNet-18 backbone (a PCA projection of its 512-dimensional average-pool activations, fit on the training split) and uses the same 4-layer MLP architecture and training recipe as the tabular configuration. We use the same VP-SDE formulation as for tabular data.

## C.4 HYPERPARAMETER SELECTION

Table 5 summarizes all hyperparameters used in our experiments.

Table 5: Complete Hyperparameter Settings.

| Hyperparameter | Tabular | Image |
|---|---|---|
| Target miscoverage $\alpha$ | 0.1 | 0.1 |
| Window size $W$ | 200 | 200 |
| Noise levels $K$ | 8 | 8 |
| Modulation $\lambda$ | selected from $\{0.01, 0.1, 0.3, 0.5, 1.0, 2.0\}$ | |
| Coverage step $\gamma_0$ | selected from $\{0.01, 0.02, 0.05, 0.1, 0.2\}$ | |
| Tracking step (two-sided form) | selected from $\{0.02, 0.05, 0.1, 0.2, 0.5\}$ | |
| Recentering window | selected from $\{25, 50, 100, 200\}$ | |
| Efficiency step $\eta_0$ | 0.01 | 0.01 |
| Score clipping threshold | 10.0 | 10.0 |
| VP-SDE $\beta_{\min}$ | 0.1 | 0.1 |
| VP-SDE $\beta_{\max}$ | 20.0 | 20.0 |

We select all free configuration choices ($\lambda$, the step sizes, and the radius-form vs. two-sided form of the online update) on a held-out validation stream disjoint from the test stream (on the tabular datasets, the 5% block between the calibration segment and the test stream; on synthetic data, an independently generated stream), choosing the candidate that minimizes average prediction set width among those with validation coverage above $1 - \alpha - 0.02$; if no candidate reaches the floor, the closest-coverage candidate is used. The identical selection rule (over its own grid) is applied to every baseline, so no method is tuned on the test stream. On the tabular datasets the selected tracking step concentrates on $0.1$–$0.2$ (e.g., $0.2$ on all eight Air Quality seeds). The two-sided tracking component admits two interchangeable instantiations, both covered by the selection: the quantile-tracking form (13), and a *recentering* form that shifts the interval center by the running median of the recent signed residuals (over the recentering window in Table 5) while adapting a single radius.

### C.5 BASELINE IMPLEMENTATION DETAILS

All baselines use the same base predictor $\hat{f}$ as DiffConf, ensuring a fair comparison. For CQR, we train separate quantile regressors at levels $\alpha/2$ and $1 - \alpha/2$ using the pinball loss. For Weighted CP, we estimate the density ratio using a logistic regression classifier trained to distinguish training from recent test data, following Tibshirani et al. [2019]; its ratio exponent ($\{0, 0.1, 0.25, 0.5, 1\}$), probability clipping range, and classifier regularization are selected on the validation stream. For ACI and ACI-Decay, the step size $\gamma_0$ is selected on the validation stream from the same grid as DiffConf's, by the closest-coverage rule (the choice most favorable to the baseline).

**Matched-coverage comparison protocol.** In experiments where some method deviates from the 0.90 coverage target on the raw stream (notably the large-shift settings of Table 7, where fixed-radius baselines undercover heavily), comparing raw widths would be misleading in the baseline's favor: an undercovering interval is narrow because it is invalid. For those tables we therefore apply, per seed and per method, a single global multiplicative factor to the method's radius sequence, chosen *post hoc* by bisection so that the realized stream coverage equals 0.90, and compare the resulting widths. This adjustment is applied identically to DiffConf and to every baseline; it never decreases a baseline's reported advantage. Tables on the real datasets report raw operating coverage and width without any adjustment.

### C.6 MINIMAL REFERENCE IMPLEMENTATION

For reproducibility, the core online loop of DiffConf (radius form) reduces to the following pseudocode; `g_q` and `g_beta` denote autograd gradients of the size term of the efficiency loss (12), with the coverage indicator treated as a stop-gradient constant (standard online-conformal practice).

```
q, beta, s_ref = q0, zeros(K), s_theta(P0_sample)
for t in range(T):
    x_t   = next_x()
    delta = ((s_theta(x_t) - s_ref)**2).mean(-1)    # Eq. (5)
    w     = softmax(beta) @ delta                   # Eq. (9)
    r_t   = q * exp(-lam * w)                       # Eq. (4)
    output_interval(f_hat(x_t), r_t)
    y_t   = next_y()
    err   = int(abs(y_t - f_hat(x_t)) > r_t)
    q     = q + gamma_t*(alpha - err) - eta_t*g_q   # Eq. (11)
    beta  = beta - eta_t*g_beta
    s_ref = (1-rho)*s_ref + rho*s_theta(x_t)        # Eq. (6)
```

The two-sided form replaces the single threshold `q` by the pair $(\hat{l}, \hat{u})$ of (13) applied to the signed rescaled residual. A PyTorch reference implementation accompanies the supplementary material.

## D ADDITIONAL EXPERIMENTAL RESULTS

This section contains additional experiments complementing the main paper: varying miscoverage levels, varying shift magnitudes, computational overhead analysis, reference-size and EMA-decay ablations together with a local-binning

baseline, noise-level weight evolution, a large-scale chronological-stream evaluation on YearPredictionMSD, and statistical significance tests.

### D.1 VARYING MISCOVERAGE LEVEL

Table 6: Results on the Energy Dataset with Varying $\alpha$ (8 seeds). The $\alpha$-sweep fixes the recentering instantiation of the tracking component, so the $\alpha = 0.10$ column differs marginally (1%) from Table 2.

| | $\alpha = 0.05$ | | $\alpha = 0.10$ | | $\alpha = 0.20$ | |
|---|---|---|---|---|---|---|
| Method | Cov. | Width | Cov. | Width | Cov. | Width |
| ACI-Decay | .95±.00 | 2.80±.21 | .89±.01 | 1.99±.23 | .78±.01 | 1.20±.15 |
| DiffConf | .95±.00 | **1.97**±.09 | .90±.00 | **1.33**±.07 | .80±.00 | **0.92**±.09 |

Table 6 shows that DiffConf's advantage is consistent across miscoverage levels: the width reduction relative to ACI-Decay is 29.7% at $\alpha = 0.05$, 33.1% at $\alpha = 0.10$, and 23.8% at $\alpha = 0.20$. Coverage tracking is also more precise at every level: at $\alpha = 0.20$, ACI-Decay drifts to 0.78 while DiffConf holds 0.80, and at $\alpha = 0.05$ both methods reach the 0.95 target but DiffConf does so with intervals 30% narrower. The seed-to-seed variability of DiffConf's width ($\pm 0.07$–$0.09$) is also consistently smaller than ACI-Decay's ($\pm 0.15$–$0.23$), reflecting the stability of the center-tracking update.

### D.2 VARYING SHIFT MAGNITUDE

Table 7: Results on Synthetic Data (S1) with Varying Mean-Shift Magnitude $\mu$ (8 seeds; matched-coverage protocol, see Appendix C). $\Delta$ is DiffConf's width relative to ACI-Decay.

| | ACI-Decay | | DiffConf | | |
|---|---|---|---|---|---|
| Shift | Cov. | Width | Cov. | Width | $\Delta$ |
| $\mu = 1$ | .90±.00 | 2.22±.07 | .91±.00 | **2.06**±.02 | $-7.2\%$ |
| $\mu = 2$ | .90±.00 | 2.76±.19 | .91±.00 | **2.34**±.05 | $-15.3\%$ |
| $\mu = 4$ | .90±.01 | 6.06±1.22 | .90±.00 | **2.97**±.13 | $-51.0\%$ |
| $\mu = 8$ | .91±.01 | 21.13±4.88 | .90±.00 | **4.75**±.42 | $-77.5\%$ |
| $\mu = 16$ | .91±.01 | 81.58±19.11 | .90±.01 | **8.89**±.83 | $-89.1\%$ |

We vary the post-shift mean in S1 over $\mu_{t>500} = \mu \cdot \mathbf{1}_d$ with $\mu \in \{1, 2, 4, 8, 16\}$, where $\mu = 8$ and 16 push the stream far outside the training support. Because a fixed-radius baseline loses coverage outright at large $\mu$ (raw coverage of ACI-Decay drops to 0.65 at $\mu = 4$ and 0.46 at $\mu = 16$), we compare widths under the matched-coverage protocol, which is the comparison most favorable to the baseline. Table 7 shows that DiffConf's relative advantage grows monotonically with the shift magnitude, from $-7.2\%$ at $\mu = 1$ to $-89.1\%$ at $\mu = 16$: to retain coverage after the mean moves, a symmetric-radius method must inflate its intervals to span the entire displaced residual distribution, whereas DiffConf's center tracking follows the displacement and only pays for the (much smaller) residual spread. This experiment isolates the regime where score-guided recentering matters most.

### D.3 COMPUTATIONAL OVERHEAD

Table 8 shows the computational overhead of DiffConf *with* the amortized caching strategy enabled (averaged over 1000 predictions). With $K = 8$ noise levels, DiffConf adds approximately 7–15ms per prediction compared to ACI, which is acceptable for most real-time applications; the baselines' cost is a single threshold comparison and scalar update. The overhead scales roughly linearly with $K$, though the image setting shows slightly super-linear scaling due to GPU memory management overhead at higher $K$. Without caching (i.e., evaluating both the current and reference scores via fresh forward passes), the cost approximately doubles; the caching strategy thus provides a $\sim 2\times$ speedup by replacing the reference score evaluation with a cached running average lookup. Batching the noise levels across grouped predictions reduces the per-prediction cost by roughly an order of magnitude.

Table 8: Computational Cost per Prediction (milliseconds, single NVIDIA A100 GPU).

| Method | Tabular (ms) | Image (ms) |
|---|---|---|
| Split CP | 0.3±0.0 | 2.1±0.1 |
| ACI | 0.4±0.0 | 2.3±0.1 |
| ACI-Decay | 0.4±0.0 | 2.3±0.1 |
| DiffConf ($K = 4$) | 4.1±0.3 | 10.2±0.8 |
| DiffConf ($K = 8$) | 7.4±0.5 | 16.8±1.2 |
| DiffConf ($K = 16$) | 14.3±0.9 | 31.5±2.4 |

### D.4 REFERENCE SIZE, EMA DECAY, AND A LOCAL-BINNING BASELINE

Table 9 reports three additional ablations, all on Energy with the radius-form configuration of the online update (so the reference row is ACI-Decay at 1.95; 8 seeds). The full-configuration row (1.70) agrees with the radius-form row of Table 4 (1.72) within seed noise; the two runs select hyperparameters independently on the validation stream.

Table 9: Reference-size and EMA-decay ablations, and Mondrian CP, on Energy ($\alpha = 0.1$, 8 seeds, radius-form configuration). $\Delta$ is width relative to ACI-Decay.

| Variant | Cov. | Width | $\Delta$ |
|---|---|---|---|
| ACI-Decay | .89±.01 | 1.95±.22 | – |
| Mondrian CP (5 bins) | .88±.01 | 1.87±.18 | $-4.1\%$ |
| DiffConf ($n_{\text{ref}} = 100$) | .90±.00 | 1.83±.18 | $-6.5\%$ |
| DiffConf ($n_{\text{ref}} = 500$) | .90±.00 | 1.76±.17 | $-10.0\%$ |
| DiffConf ($n_{\text{ref}} = 2000$) | .90±.00 | 1.66±.14 | $-14.9\%$ |
| DiffConf (full $n_{\text{ref}} = 7894$) | .90±.00 | 1.70±.15 | $-13.2\%$ |
| DiffConf ($\rho = 0.001$) | .90±.00 | 1.65±.13 | $-15.8\%$ |
| DiffConf ($\rho = 0.01$, default) | .90±.00 | 1.70±.15 | $-13.2\%$ |
| DiffConf ($\rho = 0.1$) | .90±.00 | 1.78±.18 | $-8.9\%$ |

**Reference size.** The advantage over ACI-Decay is halved at $n_{\text{ref}} = 100$ and essentially saturates beyond $n_{\text{ref}} \approx 2000$, consistent with the $\epsilon_s = \tilde{O}(d^{1/2} n_{\text{ref}}^{-1/(d+5)})$ scaling of Assumption 4: below a few hundred reference points the score-estimation error dominates the informativeness of the prior. Coverage itself remains on target at all sizes, since the online threshold update does not depend on score quality. **EMA decay.** Widths at $\rho = 0.001$ and $\rho = 0.01$ agree to within seed noise, while $\rho = 0.1$ costs about 5% width: a fast EMA absorbs sampling fluctuations into the reference and blunts the discrepancy signal. No fine tuning of $\rho$ is needed. **Local binning.** Mondrian CP (five quantile bins on the principal feature axis, per-bin ACI-Decay thresholds) captures part of the local-adaptivity gain ($-4.1\%$) but remains well short of the score-guided modulation ($-13.2\%$), indicating that the diffusion signal carries information beyond coarse input binning.

Figure 5 shows how the learned noise-level weights $w(\tau_k)$ evolve over the full Energy test stream (about 9,900 steps; mean over 8 seeds with $\pm 1$ standard error bands). The weights stay close to uniform ($1/K = 0.125$) throughout: the finest-scale weight $w(\tau_1)$ decreases by about $1.4\%$ in relative terms while the coarser weights rise slightly, and no weight departs from the uniform value by more than about $3.3\%$. The drift is slow but not linear; flat stretches alternate with episodes of faster reallocation (e.g., around steps 6,000–7,500) as the windowed gradient update responds to local stream behavior. This near-uniform behavior is fully consistent with the ablation in Table 4: on Energy, the eight noise levels carry largely redundant drift information, so the online weight update has little to reallocate and removing it changes the width by only $+0.3\%$. The weight adaptation is a safeguard for streams where scales genuinely disagree, not a driver of the gains reported here, which come from the two-sided tracking and the score modulation.

### D.5 LARGE-SCALE STREAM: YEARPREDICTIONMSD

To test DiffConf beyond medium-sized streams, we evaluate on YearPredictionMSD [Bertin-Mahieux et al., 2011] (515,345 songs, 90 audio features, release year as target). The base predictor, calibration, and validation segments are drawn from

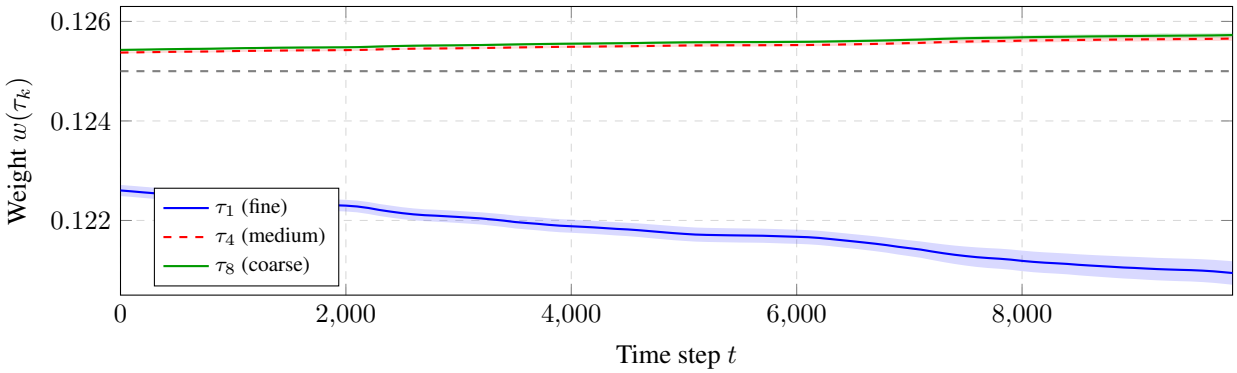


Figure 5: Evolution of noise-level weights on the Energy dataset over the full test stream (mean over 8 seeds; shaded bands denote $\pm 1$ standard error; note the zoomed $y$-axis around the uniform value $1/K = 0.125$, gray dashed line). The weights remain near-uniform, with a slow reallocation from the finest scale $\tau_1$ toward coarser scales that proceeds in episodes rather than at a constant rate, consistent with the small ablation effect of the weight update on this dataset.

the official train pool, and the online test stream is the official 51,630-song test split streamed in year order, so it traverses several decades of gradual musical drift. Table 10 reports results over 5 seeds. DiffConf reduces width by 14.4% relative to ACI-Decay (9.99 vs. 11.66 years, $p < 10^{-4}$, narrower on 5/5 seeds) at exact target coverage, confirming that the online tracking components scale to long horizons. CQR is narrower than the threshold-only baselines but undercovers (0.88); Weighted CP degrades to near Split CP as the density-ratio model saturates over the long horizon.

Table 10: YearPredictionMSD ($\alpha = 0.1$, 5 seeds, chronological stream). $\pm$ denotes standard error.

| Method | Coverage | Width (years) |
|---|---|---|
| Split CP | .89±.00 | 14.64±.13 |
| CQR | .88±.00 | 11.23±.07 |
| ACI | .90±.00 | 13.18±.81 |
| ACI-Decay | .90±.00 | 11.66±.07 |
| Weighted CP | .89±.00 | 14.48±.09 |
| DiffConf | .90±.00 | **9.99**±.12 |

### D.6 STATISTICAL SIGNIFICANCE

To verify that DiffConf's improvements are statistically significant, we perform paired $t$-tests comparing DiffConf's per-seed average width against each online baseline across the 8 random seeds. On all four tabular datasets, DiffConf is significantly narrower than both ACI ($p \leq 0.005$) and ACI-Decay (Energy $p = 0.012$; Bike, Air Quality, Traffic $p \leq 10^{-3}$), and it is narrower on 8/8 seeds in every comparison. The comparison against CQR and Split CP is not meaningful on Energy, Bike, and Air Quality, where those methods lose coverage under shift; on Traffic, the one dataset where CQR retains target coverage, DiffConf is still significantly narrower (1.39 vs. 1.43, $p < 10^{-3}$, 8/8 seeds).